Revue québécoise de linguistique

Le temps recouvré : de la pertinence d'une constitution de corpus; méthodologie et usage

Marc St-Pierre, Daniel Gosselin, Monique Lemieux et Marthe Faribault

Volume 20, numéro 2, 1991

Linguistique au Québec

URI : https://id.erudit.org/iderudit/602714ar
DOI : https://doi.org/10.7202/602714ar

Aller au sommaire du numéro

Éditeur(s)

Université du Québec à Montréal

ISSN

0710-0167 (imprimé)
1705-4591 (numérique)

Découvrir la revue

Citer cette note

St-Pierre, M., Gosselin, D., Lemieux, M. & Faribault, M. (1991). Le temps recouvré : de la pertinence d'une constitution de corpus; méthodologie et usage. *Revue québécoise de linguistique*, *20*(2), 263–279. https://doi.org/10.7202/602714ar

Résumé de l'article

Dans l'analyse, en syntaxe historique, d'un ancien état de langue, les attentes récentes de la théorie ne sont pas toujours comblées par les études et les grammaires traditionnelles de l'ancienne langue. Dans notre étude diachronique du français, nous devons donc mettre sur pied un corpus tiré de textes médiévaux afin d'accéder directement aux données qui répondent à nos questionnements théoriques. Le présent article explique les justifications pour la création de notre corpus du moyen français et pour le choix des textes qui le composent. On y présente d'abord les étapes qui permettent un traitement informatisé pratique et rapide du corpus, et ensuite les possibilités qu'offre ce traitement.

LE TEMPS RECOUVRÉ: DE LA PERTINENCE D'UNE CONSTITUTION DE CORPUS; MÉTHODOLOGIE ET USAGE*

Marc St-Pierre, Daniel Gosselin,
Monique Lemieux et Marthe Faribault

1. Introduction

L'avènement de l'informatique a profondément modifié les habitudes de travail en diachronie, comme dans tous les autres domaines de la linguistique. Choisir la voie de l'analyse de textes à l'aide de l'ordinateur, c'est cependant emprunter un chemin semé d'embûches, se lancer dans une entreprise dont les étapes doivent être soigneusement planifiées.

Cet article a donc pour but de mettre en lumière les défis méthodologiques que nous posent la constitution et l'analyse d'un corpus informatisé. Il sera organisé comme suit: les raisons qui nous ont amenés à créer un corpus (section 2); les étapes de la constitution de ce dernier (section 3); et finalement un bref aperçu de l'utilisation que nous en faisons (section 4).

2. La justification du corpus

Si l'on considère l'ensemble des données à notre disposition sur le moyen français, il peut sembler aberrant de se lancer dans une entreprise de constitution de corpus. Mentionnons parmi les travaux les plus connus les grammaires de Marchello-Nizia (1979), de Martin et Wilmet (1980), les études de Lewinsky (1948), de Nissen (1943), de Dees (1978), de Price (1973) et de Martin (1978). Plusieurs de ces études fournissent des statistiques sur divers aspects de la syntaxe

* Cette recherche est subventionnée par le CRSH (# 410-89-1409). Nous tenons aussi à remercier les autres membres de l'équipe pour leur encouragement et leur aide: Anne-Marie Benoit, Mariette Champagne, Monique Dufresne, Fernande Dupuis, Carole La Grenade et Pierre Pica.

du moyen français. Pourquoi alors ne pas nous contenter de ce qui existe ? Les raisons sont multiples:

2.1 La non exhaustivité des grammaires

Les grammaires illustrent les tendances générales d'une période de la langue; elles indiquent habituellement les phénomènes les plus courants et ceux qu'on considère comme marginaux, mais elles n'ont pas comme objectif d'étudier en profondeur un aspect en particulier. Se limiter aux grammaires générales pour développer une analyse formelle de la syntaxe du moyen français peut conduire à des généralisations hâtives que les faits de langue ne confirment pas. Sur le plan théorique, il peut être intéressant, par exemple, d'établir un lien, comme le fait Kayne (1989), entre l'omission du sujet et la séparation de «ne» et de l'adverbe négatif dans les infinitives («n'avoir point de haine»). Mais de nombreuses études empiriques, comme Moignet (1965), Skårup (1975), de Kok (1985) et Pearce (1990), démontrent que cette séparation n'était caractéristique ni de l'ancien, ni du moyen français. Le phénomène semble plutôt s'être manifesté au moment de la perte des caractéristiques de langue «*pro-drop*». Un second exemple: les grammaires[1] mentionnent les particularités de l'accord du participe passé. Nos travaux[2] ont permis de mettre en lumière l'importance de la variabilité de cet accord avec «avoir», de même que le rôle du cas dans l'accord avec l'auxiliaire «être». Seules des études sur des corpus importants peuvent mettre de tels faits en lumière. Elles peuvent également nous conduire à une révision du découpage qu'on fait des périodes de l'histoire du français, comme le souligne Benucci (1988), p. 6.

> «Une périodisation correcte de l'histoire d'une langue ne peut être faite qu'en se fondant sur l'analyse des faits linguistiques. Seulement une analyse suivie et rapprochée, dans un *corpus* linguistique suffisamment ample et couvrant entièrement l'arc temporel de l'histoire de la langue en question, des principaux phénomènes des différents niveaux linguistiques (et surtout du niveau syntaxique, le plus profond qui soit), et l'analyse de leur évolution dans un cadre théorique donné et univoque, permettra d'identifier les tournants majeurs de chaque phénomène et d'établir

1. Citons celles de Foulet (1928), de Ménard (1968), de Moignet (1979) et de Marchello-Nizia (1979), par exemple.
2. Voir Benoit et Dupuis (1990).

> par là des faisceaux de «chronoglosse», autour desquels on pourra fixer les charnières de l'histoire de la langue considérée, et déterminer ensuite le nombre des phases dont elle se compose.»

2.2 Le manque de perspective théorique

La plupart des grammaires du moyen français et des monographies à notre disposition n'adoptent pas un cadre théorique; leurs auteurs n'ont pas nécessairement le même intérêt que le nôtre, en grammaire générative, en face d'un problème donné et, par rapport aux questions qui nous intéressent, ne fournissent que des données parcellaires: ainsi l'asymétrie principales / subordonnées dans l'omission du sujet est généralement mentionnée, mais on ne peut pas savoir si l'omission du sujet dans les subordonnées est un phénomène marginal ou relativement important. Dans le corpus que nous avons étudié, la moyenne d'omission du sujet est de 31% dans les principales et de 16% dans les subordonnées. C'est donc un phénomène dont il faut tenir compte aussi en subordonnée.

2.3 La difficulté de transposition des données

Certaines monographies présentent une analyse détaillée appuyée par des statistiques, mais ces données sont rarement utilisables parce que les bases de calculs ne correspondent pas aux critères d'analyse que nous nous sommes donnés. Par exemple, Nissen (1943), dans son étude de la *Chronique de Jean d'Outremeuse*, relève 2232 cas de sujets inversés en principales, soit 42,9% parmi lesquels il inclut les incises. Et comme l'auteur ne tient pas compte des sujets omis, ce pourcentage correspond à la proportion des sujets inversés par rapport aux sujets présents. Lewinsky (1949), dans son étude de *Bérinus*, n'inclut pas les sujets nuls non plus.

Quelle que soit son importance et sa représentativité, un corpus ne remplacera jamais un locuteur natif. En diachronie des langues mortes, le corpus ne permettra jamais à l'analyste de poser un jugement de locuteur natif. Il lui permettra cependant de vérifier certaines hypothèses théoriques et, si le corpus est représentatif, de situer les moments cruciaux de l'évolution. La représentativité des corpus est une question qui devrait faire l'objet d'un débat parmi les linguistes qui s'intéressent à la théorie du changement.

Marchello-Nizia (1985) dégage les principales caractéristiques d'une langue de corpus: (i)- *C'est une langue hybride*: il est parfois difficile de savoir si l'on a affaire à la langue de l'auteur ou à celle d'un copiste; (ii)- *Elle représente un registre écrit:* en d'autres mots, même un gros corpus ne nous donne pas le portrait de la langue parlée à une période donnée du moyen français; (iii)- *C'est une langue marquée de traits dialectaux*: on peut décider ou non d'inclure cette variable dans les critères de constitution d'un corpus; (iv)- *C'est une langue en partie déterminée par la composition*: la langue des vers est parfois assez différente de la langue de la prose.

En dépit de ses limites, le corpus demeure le seul moyen de s'approprier la langue d'une époque révolue. Selon les dimensions qu'on veut donner à une recherche, plusieurs critères peuvent donc servir à l'élaboration d'un corpus; il est important de souligner que ceux que nous avons retenus se justifient par notre souci de comprendre non seulement les paramètres linguistiques d'un changement, mais aussi les facteurs externes qui en ont influencé la diffusion.

3. La constitution du corpus

Au terme de la présente phase de la recherche, nous disposerons d'une trentaine de textes couvrant trois siècles: XIV^{e}, XV^{e} et XVI^{e} siècles. Dans cette section, nous commenterons brièvement les critères ayant présidé à nos choix. Dans un second temps, nous décrirons sommairement les moyens utilisés pour informatiser ces textes.

3.1 Le choix des textes

Si le philologue se fait souvent éditeur, et dans une certaine mesure créateur de textes anciens, (Cerquiglini, 1989), le linguiste diachronicien pour sa part se limite généralement à un rôle d'utilisateur. La valeur de ses travaux repose donc, en partie du moins, sur la fiabilité des textes édités dont il aura extrait ses données de base. C'est dire que le choix des textes eux-mêmes, la détermination de la meilleure édition disponible et l'évaluation de la représentativité de chaque texte au sein du corpus général sont des opérations fondamentales pour la bonne marche de sa recherche.

Dans le cas qui nous occupe, l'ensemble du corpus a été défini avec l'aide de spécialistes de langue médiévale[3]. En premier lieu, il s'agissait de choisir des oeuvres chronologiquement bien étalées à l'intérieur de la fourchette temporelle définie, soit de 1300 à 1600 (voir le tableau 1), en répartissant autant que possible entre oeuvres majeures, comme par exemple celles de Charles d'Orléans ou de Montaigne, et oeuvres de moindre importance. Pour chaque oeuvre d'auteur connu, il a fallu ensuite établir un bref dossier historique qui fasse la synthèse des connaissances sur sa biographie (en particulier: date et lieu de naissance, lieux de séjour, langues connues, éducation reçue) et les dates et lieu de composition de l'oeuvre retenue. Les renseignements fournis par ce dossier permettent, par exemple, d'interpréter certains résultats anormaux obtenus en cours de recherche. Le cas classique est celui d'un auteur qui écrit à un âge très avancé, disons en 1500 à 80 ans, dont la langue représente vraisemblablement l'usage général de 1460, lorsqu'il avait 40 ans.

En deuxième lieu, pour chaque texte retenu, il s'agissait de déterminer quelle était la meilleure édition disponible du point de vue d'une recherche en syntaxe historique. L'analyse des diverses éditions permet de se rendre compte que la plupart des éditions de textes anciens effectuées au XIX^{e} siècle et jusqu'au début du XX^{e} siècle sont difficilement utilisables parce qu'elles reposent sur une règle de reconstruction d'un original perdu. Autrement dit, l'éditeur se permet d'inventer ce texte original perdu, à partir des manuscrits transmis. En définitive, ces éditions présentent souvent non pas la langue ancienne telle qu'elle s'écrivait, mais la langue ancienne telle que se l'imaginaient les philologues. Certaines de ces éditions vont si loin dans la reconstruction qu'il devient impossible de démêler à l'intérieur du texte ce qui est le fait de la tradition manuscrite et ce qui est le fait de l'éditeur; ces éditions sont absolument inutilisables pour les recherches en syntaxe historique. D'autres éditeurs, plus humbles dans leurs visées de reconstruction du mythique original perdu, intervenaient moins et prenaient la peine de signaler et de discuter minutieusement chaque correction effectuée; ces éditions sont déjà plus utilisables, à condition de retourner à la leçon originale lorsque cela s'avère nécessaire. Mais ce n'est qu'à partir des années '30, soit après les critiques faites à la pratique philologique par

3. Nous remercions aussi madame Christiane Marchello-Nizia avec qui nous avons pu discuter de nombreuses questions de choix de textes.

Joseph Bédier, que des éditions satisfaisantes du point de vue de la recherche en syntaxe historique commencent à paraître. Les textes sont non plus reconstruits à partir de l'étude de l'ensemble de la tradition manuscrite, mais bien restitués à partir du meilleur manuscrit conservé, qui ne sera amendé par l'éditeur que dans les seuls cas de fautes patentes, révélées par les autres manuscrits conservés de l'oeuvre ou par ce que l'on connaît de façon sûre de la langue de l'époque; quel que soit le cas, la moindre intervention de l'éditeur sera soigneusement signalée et justifiée dans les notes critiques. Les textes édités suivant cette règle sont en principe fiables pour les travaux de recherche en syntaxe historique. Cependant on peut encore leur faire un reproche: celui de considérer comme mineur, et par conséquent non digne de figurer dans les variantes textuelles, le fait de déplacer un mot dans une phrase sans que cela ne change le sens; ainsi un sujet, pronominal ou non, placé devant le verbe dans le manuscrit de base, mais placé derrière le verbe dans un manuscrit secondaire, ne sera pas noté dans l'apparat critique; or, c'est justement le genre de données que recherche le linguiste diachronicien.

Un second critère de choix de texte est le style d'écriture, selon qu'il s'agit d'un texte en vers ou en prose. Dans une perspective diachronique, la langue des vers nous informe autant que la langue de la prose. La langue des vers est certes une langue archaïsante, mais elle n'en est pas moins le reflet des structures disponibles à une époque donnée. Elle peut aussi nous apporter des indications précieuses sur des phénomènes prosodiques.

Jusqu'à maintenant, cette projection de choix de textes nous donne la répartition du tableau 1.

Tableau 1
La répartition des textes du corpus selon le temps et le style d'écriture.

PÉRIODE	TYPE	NOM DES OEUVRES
1/3 XIVe s.	PROSE	Roman d'Auberon
2/3 XIVe s.	VERS	Miracles de Notre Dame
2/3 XIVe s.	PROSE	Bérinus, vol I / La Vie de St-Augustin I
3/3 XIVe s.	PROSE	Mélusine / Chroniques (Froissart)
1/3 XVe s.	PROSE	XV joies de Mariage / Le Livre du Corps de Policie
2/3 XVe s.	VERS	Poésies (C. Orléans) / Oeuvres (Vaillant)
2/3 XVe s.	MIXTE	Oeuvres (P. Chastellain)
2/3 XVe s.	PROSE	Cent Nouvelles Nouvelles / L'Abuzé en Court
3/3 XVe s.	PROSE	Mémoires (P. Commynes) / Cleriadus et Meliadice/ La Vie de St-Augustin II
1/3 XVIe s.	VERS	Oeuvres (J. Marot) / Oeuvres (C. Marot)
2/3 XVIe s.	PROSE	Débat de Folie et d'Amour / Heptaméron des Nouvelles
3/3 XVIe s.	PROSE	Essais (M. Montaigne) / Oeuvres (R. Garnier)

Notons immédiatement que l'on retrouve deux versions de *La Vie de St-Augustin*. Il est intéressant de comparer la traduction de la *Legenda Aurea* de Jacques de Voragine faite par Jean de Vignay et la révision de Jean de Batallier[4] un siècle plus tard, cet aspect faisant présentement l'objet de l'étude de Lemieux (1990).

Nous n'avons pas cherché à inclure systématiquement un nombre égal de textes des dialectes de l'île de France, du francien, et des autres dialectes. Le corpus actuel est largement dominé par des textes écrits en francien. Les travaux de Dees

4. Brenda Dunn-Lardeau procède actuellement à l'édition de la traduction de Jean de Batallier. Nous la remercions d'avoir mis cette version à notre disposition.

(1978) ont mis en lumière l'importance de cette variable et nous comptons l'inclure dans la poursuite de nos travaux.

3.2 Les outils informatiques

Divers systèmes informatiques font partie intégrante de notre procédure de prise de corpus et de son analyse. Afin de mieux suivre les étapes de l'informatisation du corpus, le diagramme 1 schématise cette procédure.

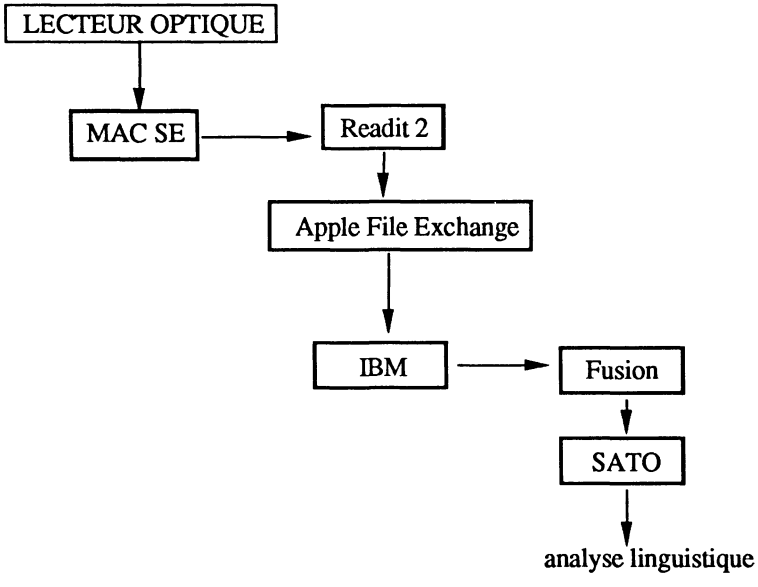


Diagramme 1
Les étapes de la constitution informatique du corpus.

Après avoir choisi l'édition d'un texte, nous en numéralisons les pages à l'aide d'un lecteur optique relié à un ordinateur Macintosh. Ces pages sont transformées en fichiers par le logiciel *Readit2*, puis transférées sur un ordinateur compatible IBM par le biais de *Apple File Exchange*.. Finalement, un logiciel maison, *Fusion*, relie ces fichiers et les soumet à un autre logiciel, *SATO* (Système d'analyse de textes par ordinateur) conçu par monsieur François Daoust du Centre d'analyse de textes

par ordinateur de l'UQAM. C'est à l'aide de *SATO* que nous codifions le corpus, ce qui nous préparera à l'analyse linguistique. Il nous faut transférer les textes sur un environnement IBM, puisque *SATO* n'est pas compatible à un environnement Macintosh.

Dans les prochaines sections, nous traiterons plus en détails de *SATO*, des propriétés symboliques et de la codification de nos textes[5].

4. L'utilisation du corpus

La codification d'un corpus se fait en fonction d'hypothèses. Dans cette section, nous rappellerons la problématique qui justifie la codification retenue. Dans un second temps nous présenterons les propriétés propres à ces codifications. Nous terminerons par les possibilités pertinentes que nous offre le logiciel *SATO* et qui nous permettent ces codifications informatiques.

4.1 La problématique

Dans le tableau 2, nous dressons une liste des textes faisant partie du corpus déjà analysé.

Au départ, nous avions l'intention de constituer un corpus qui nous fournirait une image longitudinale de l'évolution des caractéristiques du phénomène *pro-drop*, soit l'omission et l'inversion du sujet dans les propositions tensées. La plupart des études sur ces questions, (citons Franzén, 1939; Zwanenburg, 1974; Adams, 1987; Hirschbühler et Junger, 1989) soulignent l'asymétrie entre les principales et les subordonnées et les traitent comme des constructions caractéristiques des phrases déclaratives[6].

5. Pour avoir plus d'informations concernant les choix que nous avons dû prendre au sujet des logiciels utilisés nous vous encourageons à nous contacter.

6. L'omission du sujet est également possible dans les phrases interrogatives, comme dans l'exemple (i) :

(i) Pour quel besoing venites ci?
«Pour quelles raisons vîntes vous ici?»
[Passion du Palatinus, XIVe siècle, page 13]

Tableau 2
Les textes du corpus Lemieux étudiés.

DATE	NOM DE L'OEUVRE	AUTEUR	ÉDITEURS
1330	Passion du Palatinus	anonyme	G. Frank
1333-1340	La Vie de St-Augustin I	Jean de Vignay[7]	manuscrit
1350-1370	Bérinus, volume I	anonyme	Bossuat
1339-1382	Miracles de Notre Dame	anonyme	Paris & Robert
1392-1393	Mélusine	Jean d'Arras	de Dijon
1404-1407	Livre du Corps de Policie	Christine de Pizan	R.H. Lucas
1382-1410	XV Joies de Mariage	anonyme	Droz & Minard
1440-1454	Oeuvres	Pierre Chastellain	R. Deschaux
1456-1467	Cent Nouvelles Nouvelles	anonyme	F.P. Sweetser
1430-1470	Poésies[8]	Charles d'Orléans	P. Champion
1445-1470	Oeuvres	Vaillant	R. Deschaux
1476	La Vie de St-Augustin	II Jean de Batallier[9]	éd. en cours
1480	L'Abuzé en Court	inconnu[10]	R. Dupuis
1467-1483	Cleriadus et Meliadice	anonyme	G. Zink
1484-1498	Mémoires	Philippe de Commynes	J. Calmette
1548-1555	Débat de Folie et d'Amour	Louise Labé	F. Rigolot

Selon de nombreuses hypothèses, dont Jaeggli (1980), Rizzi (1982), Bouchard (1984) et Travis (1984), l'évolution du phénomène «*pro-drop*» serait soumise à des influences morphologiques: la reconnaissance des traits de la catégorie vide «pro» dépend des traits de personne sur la forme verbale. Pour vérifier cette hypothèse, le

7. Nom du traducteur.
8. La date concernant *Poésies* de Charles d'Orléans est arbitraire. On sait de ce texte qu'il provient du 2e tiers du XVe siècle.
9. Nom du traducteur.
10. Les spécialistes ne s'entendent pas à attribuer le texte au roi René d'Anjou, à Charles de Rochefort ou à un auteur anonyme.

verbe sera au centre de tous les aspects codifiés, aussi bien d'un point de vue morphologique que syntaxique. La vérification du statut de langue à verbe en deuxième position dépendra en partie de l'identification de l'élément précédant le verbe.

4.2 Les codifications linguistiques

Chaque propriété symbolique représente une propriété linguistique assignée aux verbes d'un texte[11]. Pour l'instant, chaque verbe est codifié: (i)- *selon le type de proposition auquel il appartient*: déclarative principale, subordonnée, relative, incise ou coordonnée, ou autre c'est-à-dire impérative, interrogative ou exclamative; (ii)- *selon la position ou l'absence du sujet*: absent, préverbal ou postverbal; (iii)- *selon le statut du sujet*: pronominal, nominal; (iv)- *selon la personne et le nombre du sujet*, s'il est pronominal; (v)- *selon la nature de l'élément précédant le verbe*, que l'on appelle déclencheur, si le sujet est absent ou postverbal: adverbe, objet direct, objet prépositionnel, particule *et* ou autre c'est-à-dire adjectif attribut, participe, infinitif ou proposition. Les participes passés non adjectivaux sont également codifiés.

Nous avons choisi dans cette étape de nous concentrer sur les informations indispensables à la vérification de nos hypothèses. Ainsi, parmi les propositions subordonnées, seules les relatives sujets sont parfois facilement identifiables à cause de leur nature. Une analyse plus fine doit faire appel à d'autres procédures.

4.3 L'utilisation de SATO

Le logiciel *SATO* nous permet une mise en relief de certains aspects qui nous préoccupent le plus dans le corpus. Il ne permet pas de modifier le contenu, comme un traitement de texte, par exemple. C'est un peu comme passer certains traits de crayon de couleur sur un texte, mais de manière informatique. Par exemple:

(1) Lors **ot** il telle angoisse et tel honte que en grant temps il ne **pot** mot dire.
'Alors, il eut si peur et tellement honte, que bientôt il ne put rien dire.'

Une fois le texte codifié à l'aide de *SATO*, il se présente comme suit:

11. Nous saisissons 800 verbes tensés par texte pour les fins de la codification.

```
*page=22/2/4

Lors ot*phrase=pri*decl=adv*sy=v*suj=i3 il telle angoisne et tel
honte que en grant temps il ne pot*phrase=sub*decl=na*sy=v*suj=p3
mot dire.

Et quant il ot*phrase=sub*decl=na*sy=v*suj=p3 prins*sy=pp cuer et
fu*phrase=coo revenuz*sy=pp a lui, si
s'encommença*phrase=pri*decl=adv*sy=v*suj=a3 a aler grant erre vers
l'eglise Saint Jehan, ou sa mere
estoit*phrase=sub*decl=na*sy=v*suj=p7 enterree*sy=pp, et se
print*phrase=coo la a complaindre et a dementer en disant :
<< Helas ! chaitiz, comme j'ay*phrase=aut esté*sy=v+ fol et non
sachant, quant mon mauvais cuer sentí*phrase=sub*decl=na*sy=v*suj=p7
si pou la mort de ma bonne mere ! Helas ! ores a primes
congnois*phrase=pri*decl=adv*sy=v*suj=i1 je bien l'amour qu'elle
avoit*phrase=sub*decl=na*sy=v*suj=p3 a moy et le grant bien que elle
me faisoit*phrase=sub*decl=na*sy=v*suj=p3, mais
c'est*phrase=pri*decl=na*sy=v*suj=p32 trop tart, car jamais n'y
pourray*phrase=pri*decl=adv*sy=v*suj=a1 recouvrer.
```

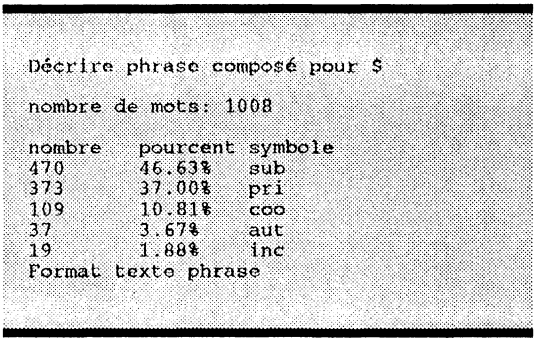
Illustration 1
Un exemple de début de page codifié dans *SATO*

Nous pouvons observer que le premier verbe de cette phrase, «ot», a été codifié comme faisant partie d'une principale («*phrase=pri»), ayant un adverbe qui le précède immédiatement («*decl=adv»), étant un verbe («*sy=v»), par opposition à un participe passé, et possédant un sujet pronominal postverbal de la troisième personne du singulier («*suj=i3»[12]).

Une fois tous les aspects d'un texte relevés par l'utilisateur, le logiciel conserve ces notes et permet la création de statistiques et de concordances. Ainsi, pour connaître le nombre relatif de phrase à verbes catégorisés dans le texte Bérinus, une

12. Il s'agit ici d'une convention par laquelle la position du sujet est identifiée par une lettre («p»=préverbal, «i»=postverbal et «a»=absent) et sa personne par un nombre (où «3»=pronominal de la troisième personne du singulier).

simple commande de *SATO* nous révèlerait les données apparaissant à l'illustration 2.



```
Décrire phrase composé pour $

nombre de mots: 1008

nombre   pourcent symbole
470      46.63%   sub
373      37.00%   pri
109      10.81%   coo
37       3.67%    aut
19       1.88%    inc
Format texte phrase
```

Illustration 2
Informations fournies par *SATO*

Lorsque l'on demande à *SATO* ce type d'informations, il peut nous le présenter directement à l'écran, permettant à l'usager une consultation rapide des données. Mais il est possible de conserver ces informations sous formes de fichiers facilement accessibles ultérieurement, soit pour la création de dossiers statistiques ou pour l'intégration dans des travaux, par le biais d'un traitement de texte.

De plus, *SATO* nous permet de faire des concordances spécifiques de propriétés ou de mots à l'intérieur d'un ouvrage du corpus. Cette fonction est fort utile pour un repérage en contexte des données déjà fournies sous forme de statistiques. Par exemple, suite aux données de l'illustration 2, si l'utilisateur veut scruter les 19 propositions incises («inc»), il peut les faire apparaître à l'écran ou dans un fichier MS-DOS. C'est une vision partielle de ces résultats de concordances que nous donne l'illustration 3.

```
Concordance libre $*phrase=inc

19 concordance(s)

          # 1 *page=22/3/8 ... *page=22/3/8/25
Seigneur >>, dist*phrase=inc Berinus, << or laissiez ce ester, car
entre mon hoste et moy nous en convendra bien.

          # 2 *page=22/3/10 ... *page=22/3/10/31
En non Dieu >>, dist*phrase=inc le bourgois, << ainsi ne finerez vous
pas a moy comme vous cuidiez, ainçois vous en convient venir >>.

          # 3 *page=22/3/16/44 ... *page=22/3/16/96
<< Sire >>, dist*phrase=inc Hanibal le prevost, << ses cinq
nefz qui sont a ce port en soient plaiges jusques a demain, et que y
mettray saisines et gardes pour ce que, s'il n'y venoit, les cinq nefz
demourront pour droit faire.
```

Illustration 3
Concordance des incises dans *Bérinus*.

Naturellement, il nous est aussi possible de faire des recherches plus fines comme d'obtenir en contexte toutes les propositions principales à sujet pronominal de la deuxième personne du singulier préverbal et dont le verbe est immédiatement suivi d'un participe passé.

Il est donc facile de voir l'utilité de *SATO* et le temps qu'il nous permet de gagner. De plus, il appert que *SATO* est suffisamment performant pour l'utilisation que nous en faisons. Néanmoins, il nous est parfois nécessaire de faire appel à d'autres logiciels (traitements de texte, chiffriers ou tableurs, bases de données) pour présenter ou gérer les données de *SATO*[13].

13. Pour des illustrations de l'exploitation de l'analyse informatisée, voir Lemieux et al. (1990), Gosselin (1990), Lemieux et Dupuis (1990) ainsi que Benoit et Dupuis (1990).

5. Conclusion

La constitution d'un corpus informatisé et la description qui en est faite à l'aide d'un logiciel d'analyse de textes est une entreprise fort complexe où chaque étape est importante, du choix des textes à la saisie par le lecteur optique à la codification suivie des multiples vérifications de cette codification. La procédure est coûteuse mais les avantages sont nombreux: repérages statistiques rapides et fiables, constitution facile de banques de données, création de concordances lexicales ou syntaxiques sans recours à la lemmatisation.

Dans cet article, nous avons communiqué les fruits de nos réflexions concernant la constitution de notre corpus. Bien sûr, nos travaux ne représentent qu'une première étape dans l'analyse de l'évolution de la langue française. Il reste encore beaucoup de travail à faire tant pour compléter le corpus, que pour raffiner notre expertise informatique et nos analyses linguistiques des données. Il est cependant déjà clair que la constitution de corpus et le développement d'instruments d'analyses tels que ceux dans lesquels nous sommes engagés permettent de mettre à la disposition des chercheurs des informations indispensables pour une étude approfondie des langues mortes, informations qu'il n'était normalement pas possible de rassembler autrefois.

Marc St-Pierre, Daniel Gosselin,
Monique Lemieux et Marthe Faribault
Université du Québec à Montréal

Références

ADAMS, M. (1987) *Old French, Null Subjects and Verb Second Phenomena*, thèse de doctorat, Université de Californie.

BENOIT, A.M. et F. Dupuis (1990) *L'accord du participe passé: Variation et configuration du VP*, conférence de NWAVE XVII, octobre 1988, manuscrit, Montréal, UQAM.

BENUCCI, F. (1988) *Les constructions modales du français des Serments de Strasbourg à nos jours: une analyse syntaxique*, thèse de doctorat, Université de Padova.

BOUCHARD, D. (1984) *On the Content of Empty Categories*, Dordrecht, Foris.

CERQUIGLINI, B. (1989) *Éloge de la variante: histoire de la philologie*, Paris, Éditions du Seuil.

DE KOK, A. (1985) *La place du pronom personnel régime conjoint en français, une étude diachronique*, Amsterdam, Rodopi.

DEES, A. (1978) «Variations temporelles et spatiales de l'ordre des mots en ancien et moyen français» dans Wilmet M. (éd.) *Sémantique lexicale et sémantique grammaticale en moyen français*, Actes du colloque, V.U.B. Centrum voor Tallen Literatuurwetenschap.

FOULET, L. (1928) *Petite syntaxe de l'ancien français*, Paris, Champion.

FRANZÉN, T. (1939) *La syntaxe des pronoms personnels sujets en ancien français*, Almqvist, Uppsala.

GOSSELIN, D. (1990) *Les sujets dits «inversés» en moyen français*, mémoire de maîtrise, Montréal, UQAM.

HIRSCHBÜHLER P. et M.O. Junger (1989) «Sujets nuls et pronominaux dans l'histoire du français: remarques», *Revue québécoise de linguistique théorique et appliquée* 8.

JAEGGLI, O.A. (1980) *On some Phonologically-Null Elements in Syntax*, thèse de doctorat, Cambridge, MIT.

KAYNE, R.S. (1989) «Null Subjects and Clitic Climbing», dans O. Jaeggli et K.S. Safir (éds) *The Null Subject Parameter*, Amsterdam, Reidel.

LEMIEUX, M. (1990) «L'omission et la postposition du sujet d'après 'La vie de Saint Augustin' dans la 'Légende dorée'», manuscrit, Montréal, UQAM, à paraître dans *Marche Romane*.

LEMIEUX, M. et al. (1990) «Variation paramétrique: l'expression du sujet en moyen français», manuscrit, à paraître dans *Language Change and Variation,* Montréal, UQAM.

LEMIEUX, M. et F. Dupuis (1990) *Arguments for an IP Analysis of the V2 Phenomena in Old and Middle French,* conférence au *First Generative Diachronic Syntax Conference* (Université de York, avril 1990), manuscrit, Montréal, UQAM.

MARCHELLO-NIZIA, C. (1979) *Histoire de la langue française aux XIV^e^ et XV^e^ siècles,* Paris, Bordas.

MARCHELLO-NIZIA, C. (1985) «Mélanges: Question de méthode», *Romania* 106.

MARTIN, R. (1978) «L'ordre des mots dans le Jehan de Saintré» dans Wilmet M. (éd.) *Sémantique lexicale et sémantique grammaticale en moyen français,* Acte du colloque, V.U.B. Centrum voor Tallen Literatuurwetenschap.

MARTIN, R. et M. Wilmet (1980) *Manuel du français du moyen âge, syntaxe du moyen français,* Bordeaux, Sobodi.

MÉNARD, P. (1968) *Manuel d'ancien français, 3. Syntaxe,* Bordeaux, Sobodi.

MOIGNET, G. (1965) *Le pronom personnel français. Essai de psycho-systématique historique,* Paris, Klincksieck.

MOIGNET, G. (1979) *Grammaire de l'ancien français,* Paris, Klincksieck.

NISSEN, H. (1943) *L'ordre des mots dans la Chronique de Jean d'Outremeuse,* Almquist, Uppsala.

PEARCE, E. (1990) *Parameters in Old French Syntax Infinitival Complements,* Dordrecht, Kluwer Academic Publishers.

PRICE, G. (1973) «Sur le pronom personnel sujet postposé en ancien français», *Revue Romane 8.*

RIZZI, L. (1982) *Issues in Italian Syntax,* Dordrecht, Foris.

SKÅRUP, P. (1975) «Les premières zones de la proposition en ancien français», *Revue Romane* 6.

TRAVIS, L. (1984) *Parameters and Effects of Word Order Variation,* thèse de doctorat, Cambridge, MIT.

ZWANENBURG, W. (1974) «Perte de la flexion nominale et fixation de l'ordre des mots en français médiéval» dans *XIV Congresso Internazionale di Linguistica e Filologia Romanze Atti,* III